Hi Friends,

**Abstract:**

Finding the hidden information and create a machine learning model is the goal of this project. As part of my analysis I took the credit card of a bank, where the a manager at the bank is disturbed with more and more customers leaving their credit card services. They would really appreciate if one could predict for them who is going to get churned so they can proactively go to the customer to provide them better services and turn customers' decisions in the opposite direction. End of this project I will create a model where it can predict the customers who is gonna get churned.

**About dataset:**

The data has been pulled from Kaggle Link below

https://www.kaggle.com/sakshigoyal7/credit-card-customers

**Context:**

A manager at the bank is disturbed with more and more customers leaving their credit card services. I would like to predict for them who is going to get churned so they can proactively go to the customer to provide them better services and turn customers' decisions in the opposite direction.

**Content:**

**Categorical Features**

• Attrition\_Flag (1: Existing Customer, 0: Attrited Customer): The Customer leave or not

• Gender (1: Male, 0: Female)

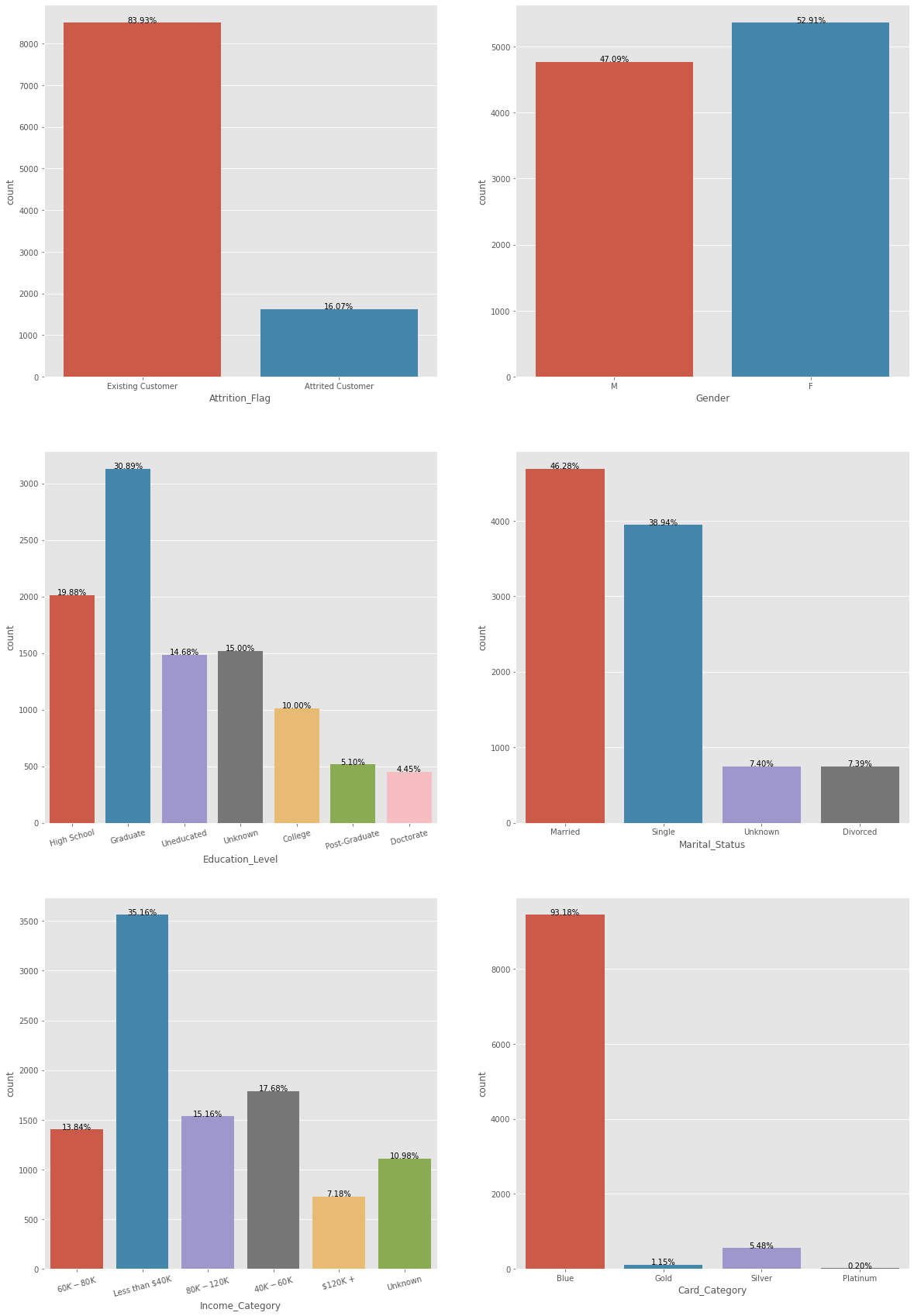
• Education\_Level (Graduate , High School, Unknown, Uneducated, College, Post-Graduate, Doctorate)

• Marital\_Status (Married, Single, Unknown, Divorced)

• Income\_Category (Less than 40K, 40K - 60K, 80K - 120K, 60K - 80K, Unknown, 120K +) in dollar

• Card\_Category (Blue, Silver, Gold, Platinum)

**1) Count plot for all Categorial Features**



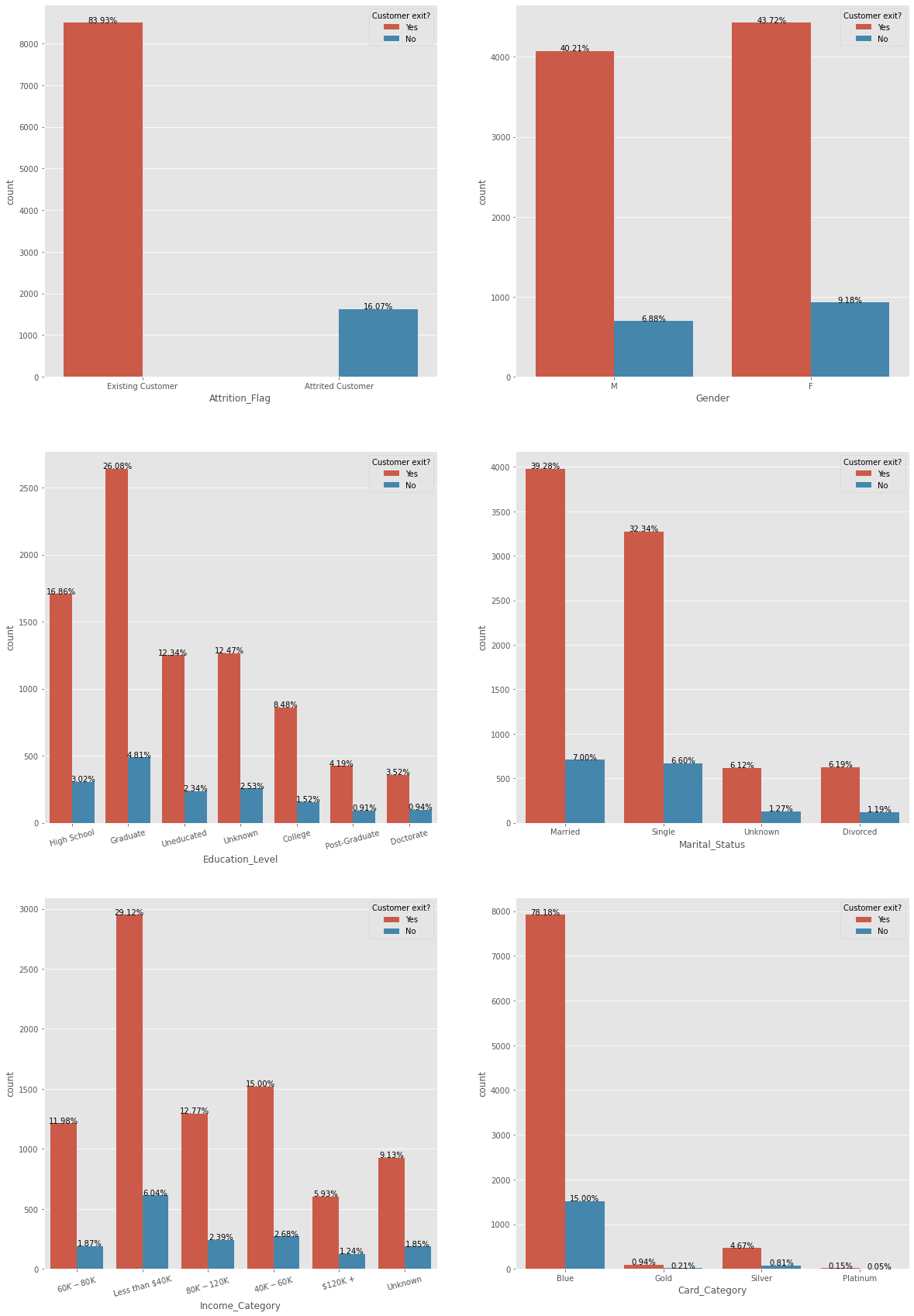
**Observations**

• We can see that the dataset is not equally distribute according to Attrition\_Flag. We have samples which are mostly Existing.

• We can say that if education level is improved, using the credit card is decreasing.

• Generally people use blue card, it's must be correlated with income.

2) Count plot for all Numerical Features



**Numerical Features**

• Customer\_Age: Customer's Age in Years

• Dependent\_count: Number of dependents

• Months\_on\_book: Period of relationship with bank

• Total\_Relationship\_Count: Total no. of products held by the customer

• Months\_Inactive\_12\_mon: No. of months inactive in the last 12 months

• Contacts\_Count\_12\_mon: No. of Contacts in the last 12 months

• Credit\_Limit: Credit Limit on the Credit Card

• Total\_Revolving\_Bal: Total Revolving Balance on the Credit Card

• Avg\_Open\_To\_Buy: Open to Buy Credit Line (Average of last 12 months)

• Total\_Amt\_Chng\_Q4\_Q1: Change in Transaction Amount (Q4 over Q1)

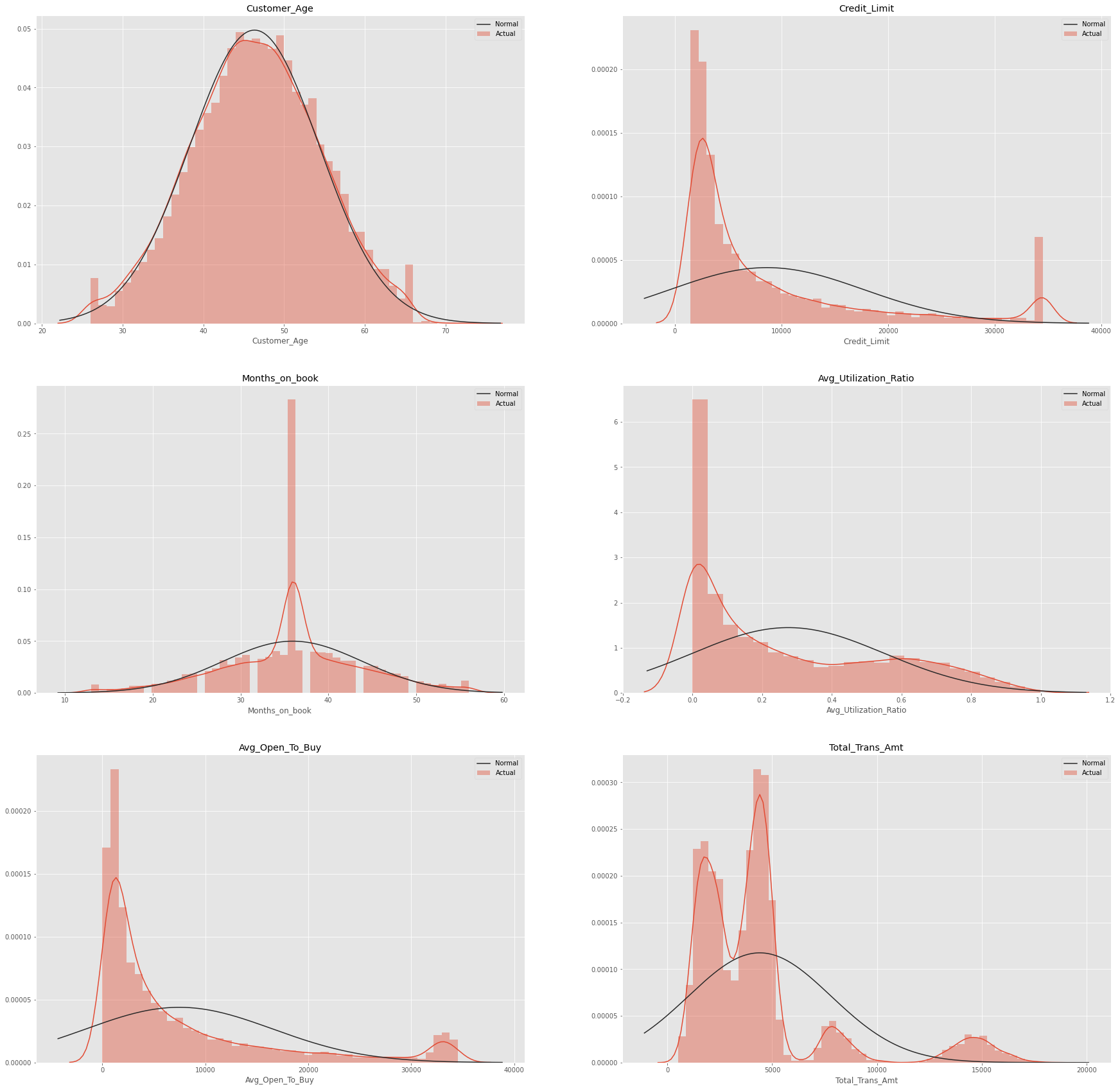
• Total\_Trans\_Amt: Total Transaction Amount (Last 12 months)

• Total\_Trans\_Ct: Total Transaction Count (Last 12 months)

• Total\_Ct\_Chng\_Q4\_Q1: Change in Transaction Count (Q4 over Q1)

• Avg\_Utilization\_Ratio: Average Card Utilization Ratio

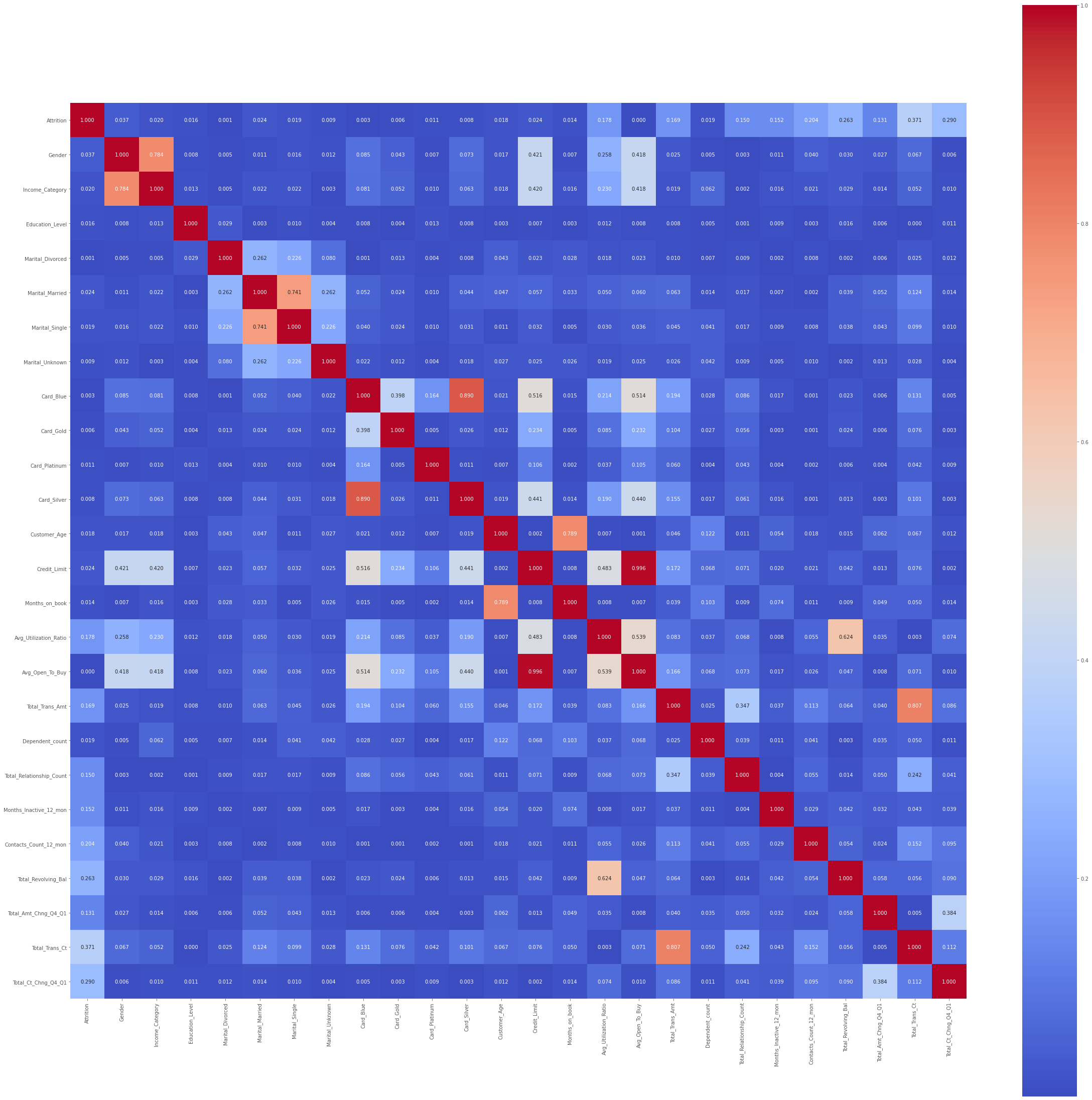
**3) Histogram for all numeric values**



**Part 2**

Before we start the Feature selection and ML algorithm. We should convert the works on numeric values. That's why we should transform Object, Category, etc. values to numeric values.

I used correlation heat map method for feature selection.



From the above correlation heat map below are the list of attributes I am going to use for ML.

• Gender

• Income\_Category

• Marital\_Married

• Marital\_Single

• Card\_Blue

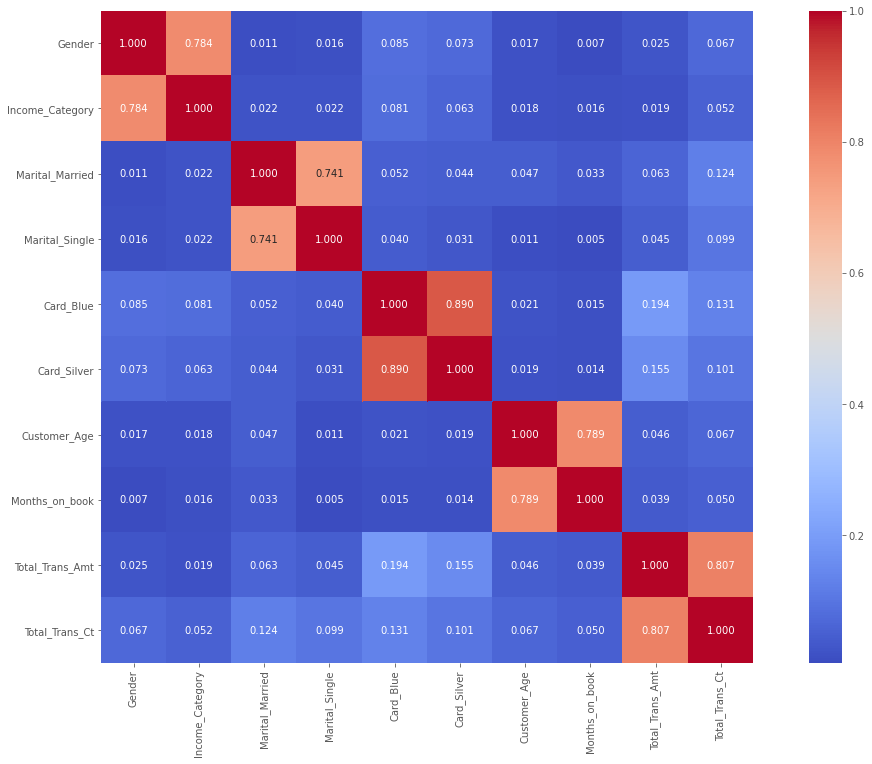
• Card\_Silver

• Customer\_Age

• Months\_on\_book

• Total\_Trans\_Amt

• Total\_Trans\_Ct

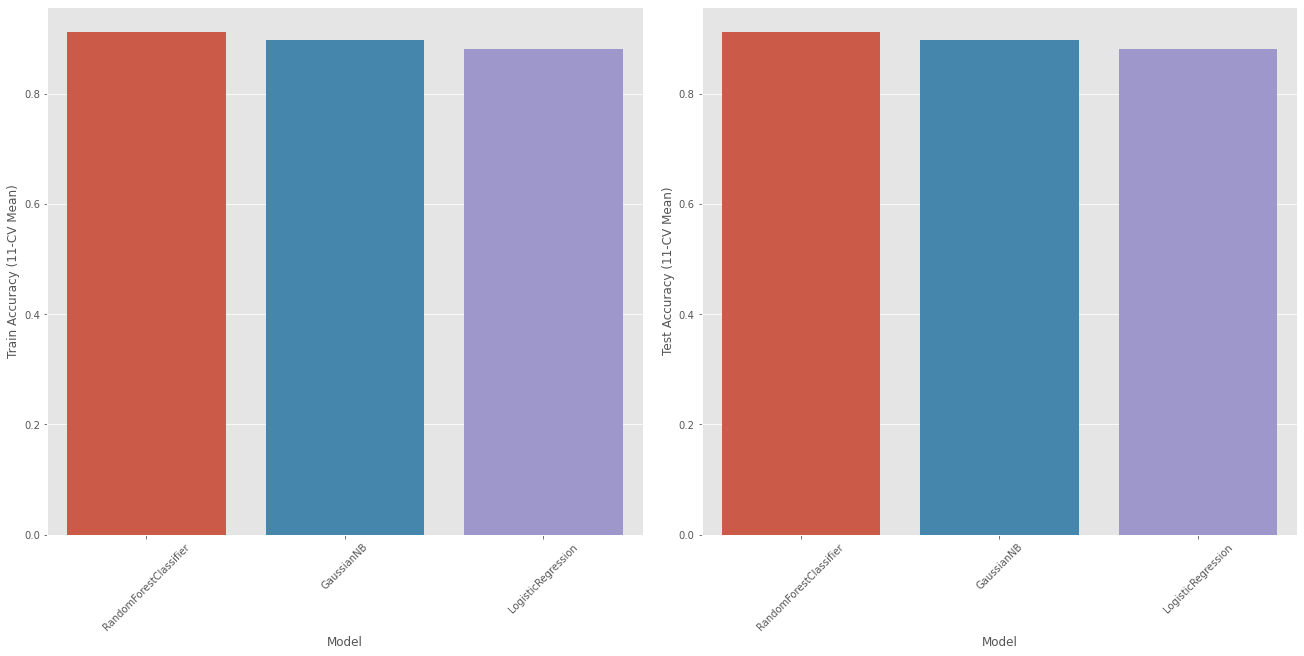


**Part 3**

The part 3, I performed model selection and model evaluation. Refer below results.

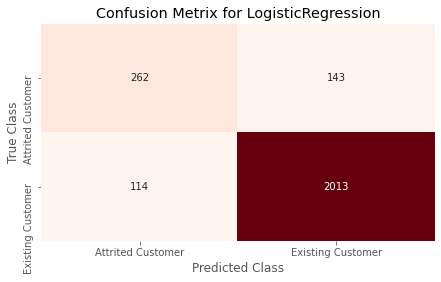
I selected LogisticRegression, RandomForestClassifier and GaussianNB models and compared the accuracy of each model for Train dataset and Test dataset and produces comparison results.

From the bellow graphs the RenadomForestClassifer has highest accuracy and Logistic Regression has lowest accuracy. In next ween I will represent the feature selection by each module.



**Part 4**

As suggested by professor and colleagues. I have made added the visualizing a classifier performance for LogisticRegression, RandomForestClassifier and GaussianNB.



**My observation says**

The LogisticRegression model predicted the 114 existing customer as attrited customers and 143 attrited customers as existing customer

The RandomForestClassifier model predicted the 24 existing customer as attrited customers and 215 attrited customers as existing customer

The GaussianNB model predicted the 93 existing customer as attrited customers and 224 attrited customers as existing customer

As per accuracy the RandomForestClassifier should perform best but LogisticRegression model is performing better that other models. It could be due to feture selections which I did.

# Conclusion

As part of this analysis I used multiple models like LogisticRegression, RandomForestClassifier and GaussianNB. Overall Random Forest Classifier seems more accurate. It is also more accurate in Existing customer findings and not recommended to find the attrited customers. However, Logistic Regression has more accuracy in terms of True or Recommended Scenarios.

Definitely, I don't want to say my model is 100% accurate. I feel as a beginner model is really performing well. Definitely there is scope of improvement and best practices to get better results. I feel there is scope of improvement of feature selection. This could make model more tuned.

Definitely we can use this model to reduce the attrited customers and save the credit card company problem.

Reference -

Applied Text Analysis with Python, Benjamin Bengfort, Rebecca Bilbro & Tony Ojeda

Machine Learning with Python Cookbook, Chris Albon